

CLAIMS

What is claimed is:

1. A method of identifying a biological activity of a compound of interest, comprising:  
providing a plurality of gene expression datasets associated with a first class of  
5 compounds having a first biological activity;  
providing a plurality of gene expression datasets associated with a second class of  
compounds having a second biological activity;  
deriving a linear classification rule based on said plurality of gene expression  
10 datasets; and  
applying said linear classification rule to a set of gene expression levels associated  
with said compound of interest thereby determining whether said compound of interest has  
said first biological activity or said second biological activity.
2. The method of claim 1, wherein each dataset comprising a set of gene expression  
15 levels and a set of gene expression intervals.
3. The method of claim 1, wherein deriving said linear classification rule includes  
deriving a linear classification function.
- 20 4. The method of claim 3, wherein deriving said linear classification function includes  
reducing a value of a loss function associated with said plurality of gene expression  
datasets.
5. The method of claim 4, wherein reducing said value of said loss function includes  
25 reducing a worse-case value of said loss function.
6. The method of claim 3, wherein deriving said linear classification function includes  
identifying a set of classifiers that minimize a value of a loss function associated with said  
plurality of gene expression datasets.
- 30 7. The method of claim 6, wherein said loss function is associated with one of a  
support vector machine, logistic regression, and minimax probability machine.

8. A method of identifying a biological state of a biological sample, comprising:  
providing a plurality of gene expression datasets, each gene expression dataset of  
said plurality of gene expression datasets including a set of gene expression levels and a set  
of gene expression intervals, said plurality of gene expression datasets including a first  
5 plurality of gene expression datasets associated with a first biological state and a second  
plurality of gene expression datasets associated with a second biological state;  
deriving a linear classification rule based on said plurality of gene expression  
datasets; and  
applying said linear classification rule to a set of gene expression levels associated  
10 with said biological sample to identify a biological state of said biological sample as one of  
said first biological state and said second biological state.
9. The method of claim 8, wherein said first biological state and said second biological state  
correspond to a normal condition and a disease condition, respectively.
- 15 10. The method of claim 8, wherein deriving said linear classification rule includes  
deriving a linear classification function.
11. The method of claim 10, wherein deriving said linear classification function includes  
20 reducing a value of a loss function associated with said plurality of gene expression  
datasets.
12. The method of claim 11, wherein reducing said value of said loss function includes  
reducing a worse-case value of said loss function.
- 25 13. The method of claim 10, wherein deriving said linear classification function includes  
identifying a set of classifiers that minimize a value of a loss function associated with said  
plurality of gene expression datasets.
- 30 14. The method of claim 13, wherein said loss function is associated with one of a  
support vector machine, logistic regression, and minimax probability machine.
15. A method for classifying a test gene expression dataset comprising:

- providing a reference gene expression dataset;  
deriving a linear classification rule by reducing the value of a loss function  
associated with said reference gene expression dataset; and  
applying said linear classification rule to a test gene expression dataset thereby  
5 determining the classification of the test gene expression dataset.
16. The method of claim 15 wherein the reference gene expression dataset is a  
chemogenomic dataset based on *in vivo* compound treatments.
- 10 17. The method of claim 15 wherein the type of loss function is selected from the group  
consisting of support vector machine, logistic regression, and minimax probability machine.
18. A computer program product for classifying a test gene expression dataset  
comprising:  
15 computer code for querying a reference gene expression dataset;  
computer code for deriving a linear classification rule by reducing the value of a loss  
function associated with said reference gene expression dataset;  
computer code for applying said linear classification rule to a test gene expression  
dataset and thereby determining the classification of the test gene expression dataset; and  
20 computer code for outputting the test dataset classification to the user.
19. The computer code product of claim 18 wherein the type of loss function is selected  
from the group consisting of support vector machine, logistic regression, and minimax  
probability machine.

25